UNIVERSITY
OF TECHNOLOGY
SYDNEY

Self-Attention Network for Text Representation Learning

by Tao SHEN

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Dist. Prof. Chengqi ZHANG, Prof. Guodong LONG and Dr. Jing JIANG

University of Technology Sydney
Faculty of Engineering and Information Technology

March 2021

Certificate of Original Authorship Template

Graduate research students are required to make a declaration of original authorship when they submit the thesis for examination and in the final bound copies. Please note, the Research Training Program (RTP) statement is for all students. The Certificate of Original Authorship must be placed within the thesis, immediately after the thesis title page.

Required wording for the certificate of original authorship

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Tao SHEN*, declare that this thesis, is submitted in fulfilment of the requirements for the award of *Doctor of Philosophy*, in the *School of Computer Science, Faculty of Engineering and Information Technology* at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.
**If applicable, the above statement must be replaced with the collaborative doctoral degree statement (see below).*

**If applicable, the Indigenous Cultural and Intellectual Property (ICIP) statement must be added (see below).*

This research is supported by the Australian Government Research Training Program.

Signature: Production Note: Signature removed prior to publication.

Date: 27 March 2021

Collaborative doctoral research degree statement

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of the requirements for a degree at any other academic institution except as fully acknowledged within the text. This thesis is the result of a Collaborative Doctoral Research Degree program with *[insert collaborative partner institution]*.

Indigenous Cultural and Intellectual Property (ICIP) statement

This thesis includes Indigenous Cultural and Intellectual Property (ICIP) belonging to *[insert relevant language, tribal or nation group(s) or communities]*, custodians or traditional owners. Where I have used ICIP, I have followed the relevant protocols and consulted with appropriate Indigenous people/communities about its inclusion in my thesis. ICIP rights are Indigenous heritage and will always remain with these groups. To use, adapt or reference the ICIP contained in this work, you will need to consult with the relevant Indigenous groups and follow cultural protocols.

ABSTRACT

Self-Attention Network for Text Representation Learning

by

Tao Shen

This research studies the effectiveness and efficiency of self-attention mechanisms for text representation learning in a deep learning-based natural language processing literature. We focus on developing novel self-attention networks to capture semantic and syntactic knowledge underlying natural language texts, and thus benefit a wide range of downstream natural language understanding tasks, which is followed by improving the networks with external relational, structured, factoid, and commonsense information in knowledge graphs.

In the last decade, recurrent neural networks and convolutional neural networks are widely used to produce context-aware representations for natural language text: the former can capture long-range dependency but is hard to parallelize and not time-efficient; the latter focuses on local dependency but does not perform well on some tasks. Attention mechanisms, especially self-attention mechanisms, have recently attracted tremendous interest from both academia and industry, due to their light-weight structures, parallelizable computations, outstanding performance on a broad spectrum of natural language processing tasks. We first propose a novel attention mechanism in which the attention between elements from input sequence(s) is directional and multi-dimensional (i.e., feature-wise). Compared to previous works, the proposed attention mechanism is able to capture the subtle difference in contexts and thus alleviate the ambiguity or polysemy problem. Based solely on the proposed attention, we present a light-weight neural model, directional self-attention network, to learn both token- and sentence-level context-aware representations, with high efficiency and competitive performance. Furthermore, we improve the proposed

network along with several directions: First, we extend the self-attention to a hierarchical structure to capture local and global dependencies for memory efficiency. Second, we introduce hard attention to the self-attention mechanism for mutual benefits of soft and hard attention mechanisms. Third, we capture both pairwise and global dependencies by a novel compatibility function composed of dot-product and additive attentions.

Then, this research conducts extensive experiments on benchmark tasks to verify the effectiveness of the proposed self-attention networks from both quantitative and qualitative perspective. The benchmark tasks, including natural language inference, sentiment analysis, semantic role labeling, are able to comprehensively estimate models' capability of capturing both semantic and syntactic information underlying natural language texts. The empirical results show the proposed models empirically achieve state-of-the-art performance on a wide range of natural language understanding tasks, and are as fast and as memory-efficient as convolutional models.

Lastly, although self-attention networks, even if those initialized by a pre-trained language model, learn powerful contextualized representations and achieve state-of-the-art performance, open questions still remain about what these models have learned and improvements can be made along with several directions. One such direction is when downstream task performance depends on relational knowledge – the kind stored in knowledge graphs. Therefore, we explore incorporation of self-attention networks and human-curated knowledge graphs because such knowledge can improve a self-attention network by either conducting symbolic reasoning over knowledge graphs to derive targeted results or embedding the relational information into neural networks to boost representation learning. We study several potential approaches in three knowledge graph-related scenarios in a natural language processing literature, i.e., knowledge-based question answering, knowledge base completion and commonsense reasoning. Experiments conducted on knowledge graph-related benchmarks show the effectiveness of our proposed models.

Key Words. Natural Language Processing; Deep Learning; Representation Learning; Contextualization Embedding; Self-Attention Network; Sentence Encoding; Question Answering; Pre-trained Language Model; Knowledge Graph

Directed by Dist. Prof. Chengqi Zhang, Prof. Guodong Long and Dr. Jing Jiang, School of Computer Science, Faculty of Engineering and Information Technology

Acknowledgements

First of all, many thanks to my principal supervisor Dist. Prof. Chengqi Zhang, and co-supervisor Prof. Guodong Long and Dr. Jing Jiang. And many thanks also to my co-worker, Tianyi Zhou, from University of Washington. During chasing the degree of Ph.D, they gave me careful guidance and kind care in my study, research, and life. Their profound knowledge, rigorous academic attitude, and tireless study spirit made me admire, and left an indelible impression on me, which greatly influenced and inspired me. I will bear their edification in my mind, and these will definitely benefit me for my whole life. I also greatly appreciate the research centre, Australian AI Institute at University of Technology Sydney, which provides a great platform to support my research works.

I also want to thank my colleagues and group mates, Peng Zhang, Lei Sang, Mengyao Li, Qian Liu, Lu Liu, ..., for their helps in all aspects of model implementing and thesis writing. The gave me tremendous help with the questions I raised and put forward many valuable suggestions for my codes and thesis, which benefit me a lot. I also appreciate the helps and inspirations from my mentors, Xiubo Geng, Tao Qin, Daxin Jiang, Yi Mao, Pengcheng He, and Weizhu Chen, during internships.

Lastly, I am deeply grateful to my parents, Kefa Shen and Yonghua Chen. Every progress I have made is condensed by their selfless dedication, support, and encouragement. Thanks to every teacher who has taught me in the growth of the patient teaching; thanks to all my friends who mutually helped to grow up together and the friendship deserves my cherish. These lovely people will continue to encourage me to move forward for more solid and comprehensive works in the future.

Tao Shen

Sydney, Australia, 2021

List of Publications

Published Conference Paper:

C-1 **Tao Shen**, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, Chengqi Zhang: DiSAN: Directional Self-Attention Network for RNN/CNN-Free Language Understanding. AAAI 2018: 5446-5455. (CORE A*)

C-2 **Tao Shen**, Tianyi Zhou, Guodong Long, Jing Jiang, Sen Wang, Chengqi Zhang: Reinforced Self-Attention Network: a Hybrid of Hard and Soft Attention for Sequence Modeling. IJCAI 2018: 4345-4352. (CORE A*)

C-3 **Tao Shen**, Tianyi Zhou, Guodong Long, Jing Jiang, Chengqi Zhang: Bi-Directional Block Self-Attention for Fast and Memory-Efficient Sequence Modeling. ICLR (Poster) 2018: 1-18. (Top conference in deep learning)

C-4 **Tao Shen**, Tianyi Zhou, Guodong Long, Jing Jiang, Chengqi Zhang: Tensorized Self-Attention: Efficiently Modeling Pairwise and Global Dependencies Together. NAACL-HLT (1) 2019: 1256-1266. (CORE A)

C-5 **Tao Shen**, Xiubo Geng, Tao Qin, Daya Guo, Duyu Tang, Nan Duan, Guodong Long, Daxin Jiang: Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base. EMNLP/IJCNLP (1) 2019: 2442-2451. (CORE A)

C-6 **Tao Shen**, Xiubo Geng, Guodong Long, Jing Jiang, Chengqi Zhang, Daxin Jiang: Effective Search of Logical Forms for Weakly Supervised Knowledge-Based Question Answering. IJCAI 2020: 2227-2233. (CORE A*)

C-7 Xueping Peng, Guodong Long, **Tao Shen**, Sen Wang, Jing Jiang, Michael Blumenstein: Temporal Self-Attention Network for Medical Concept Embedding. ICDM 2019: 498-507. (CORE A*)

- C-8 Yang Li, Guodong Long, **Tao Shen**, Tianyi Zhou, Lina Yao, Huan Huo, Jing Jiang: Self-Attention Enhanced Selective Gate with Entity-Aware Embedding for Distantly Supervised Relation Extraction. AAAI 2020: 8269-8276. (CORE A*)
- C-9 Xueping Peng, Guodong Long, **Tao Shen**, Sen Wang, Jing Jiang: Self-Attention Enhanced Patient Journey Understanding in Healthcare System. ECML 2020. (CORE A)
- C-10 Xueping Peng, Guodong Long, **Tao Shen**, Sen Wang, Jing Jiang, and Chengqi Zhang: BiteNet: Bidirectional Temporal Encoder Network to Predict Medical Outcomes on Electronic Health Records. ICDM 2020. (CORE A*)
- C-11 **Tao Shen**, Yi Mao, Pengcheng He, Guodong Long, Adam Trischler, Weizhu Chen: Exploiting Structured Knowledge in Text via Graph-Guided Representation Learning. EMNLP 2020. (CORE A)
- C-12 Yang Li, **Tao Shen**, Guodong Long, Jing Jiang, Tianyi Zhou and Chengqi Zhang: Improving Long-Tail Relation Extraction with Collaborating Relation-Augmented Attention. COLING 2020. (CORE A)
- C-13 Hao Huang, Guodong Long, **Tao Shen**, Jing Jiang and Chengqi Zhang: RatE: Relation-Adaptive Translating Embedding for Knowledge Graph Completion. COLING 2020. (CORE A)
- C-14 Bo Wang, **Tao Shen**, Guodong Long, Tianyi Zhou, Ying Wang and Yi Chang: Structure-Augmented Text Representation Learning for Efficient Knowledge Graph Completion. WWW 2021. (CORE A*)

Contents

Certificate ii

Abstract iii

Acknowledgments vi

List of Publications vii

List of Figures xiii

Abbreviation xv

Notation xvii

1 Introduction 1

1.1 Background 1

1.1.1 Natural Language Processing and Deep Learning 1

1.1.2 Text Representation Learning 3

1.1.3 Era of Pre-Trained Self-Attention Network 9

1.2 Research Objectives 12

1.3 Thesis Organization 13

2 Literature Review 14

2.1 Traditional Representation Learning 14

2.1.1 Word Embedding 15

2.1.2 Recurrent Neural Network 16

2.1.3 Convolutional Neural Network 18

2.2 Attention based Representation Learning 19

2.2.1 Attention Mechanism 19

2.2.2 Self-Attention Mechanism 22

2.2.3 Self-Attention Network 25

2.2.4 Pre-Trained Self-Attention Network 30

2.2.5 Pre-Trained Network with Knowledge Graph 32

2.3 Taxonomy of Natural Language Processing Tasks 35

3 Multi-Dimensional Attention Network 39

3.1 Introduction 39

3.2 Multi-Dimensional Attention Mechanism 41

3.3 Directional Self-Attention Network 44

3.3.1 Directional Self-Attention mechanism 45

3.3.2 Directional Self-Attention Network 49

3.3.3 Visualization of Directional Attention 50

3.4 Block Self-Attention Network 52

3.4.1 Masked Block Self-Attention Mechanism 53

3.4.2 Bi-directional Block Self-Attention Network 55

3.5 Reinforced Self-Attention Mechanism 56

3.5.1 Reinforced Sequence Sampling (RSS) 58

3.5.2 Reinforced Self-Attention (ReSA) 59

3.5.3 Model Training 62

3.5.4 Visualization of Hybrid Attention 63

3.6 Multi-Mask Tensorized Self-Attention Network 65

3.6.1 Tensorized Self-Attention (TSA) Mechanism 65

3.6.2 Multi-Mask Tensorized Self-Attention (MTSA) Mechanism . . 68
3.6.3 A Memory- and Computation-Optimized Scheme for MTSA . 69
3.7 Performance Studies 71
3.7.1 Natural Language Inference 72
3.7.2 Semantic Role Labeling 74
3.7.3 Sentence Classifications 75
3.8 Conclusion 77

4 Self-Attention Network with Knowledge Graph 79

4.1 Knowledge-based Question Answering 79
4.1.1 Introduction and Task Definition 79
4.1.2 Multi-Task Learning based Approach 84
4.1.3 Experimental Study 90
4.2 Knowledge Graph Completion 98
4.2.1 Introduction and Background 98
4.2.2 Semantic Triple Encoder for Graph 100
4.3 Experiment 107
4.3.1 Evaluations on Link Prediction 108
4.3.2 Comparison with KG-BERT Baseline 110
4.3.3 Generalization to Open-Set Prediction 110
4.3.4 Ablation Study 112
4.4 Representation learning with Knowledge Graph 113
4.4.1 Introduction and Background 113
4.4.2 Graph-Guided Representation Learning 116
4.4.3 Experimental Study 121

4.5 Conclusion 132

5 Summary and Discussion 133

5.1 Conclusion 133

5.2 Discussion and Future Work 135

List of Figures

2.1 Basic recurrent neural network (RNN) and its unroll status. 17

2.2 One-dimensional Convolutional Neural Network (1D-CNN). Copied from (Kim, 2014) 18

2.3 The Transformer, a sequence-to-sequence self-attention network. Copied from (Vaswani et al., 2017). 26

2.4 Bidirectional Encoder Representations from Transformers (BERT) in pre-training on two self-supervised tasks (left) and fine-tuning on various natural language processing tasks (right). Copied from (Devlin et al., 2019). 30

3.1 Two pairs of attention probability comparison of same word in difference sentence contexts. 43

3.2 Directional self-attention (DiSA) mechanism. Here, we use $l_{i,j}$ to denote $f(h_i, h_j)$ in Eq. (3.5). 46

3.3 Masked Self-Attention Mechanism 47

3.4 Three positional masks: (a) is the diag-disabled mask $\boldsymbol{M}^{diag}$; (b) and (c) are forward mask $\boldsymbol{M}^{fw}$ and backward mask $\boldsymbol{M}^{bw}$, respectively. 47

3.5 Directional self-attention network (DiSAN) 49

3.6 Attention probability in forward/backward DiSA blocks for the two example sentences. 51

3.7 Fusion Gate $\boldsymbol{F}$ in forward/backward DiSA blocks. 52

3.8 Masked block self-attention (mBloSA) mechanism. 54

3.9 Bi-directional block self-attention network (Bi-BloSAN) for sequence embedding. 56

3.10 Reinforced self-attention (ReSA) model. $f_{i,j}$ denotes the alignment score obtained from $f(\boldsymbol{x}_i, \boldsymbol{x}_j)$. 60

3.11 Attention probabilities of soft self-attention in ReSA. The tokens aligned in horizontal axis are heads, and the tokens aligned in vertical axis are dependents. 64

3.12 Tensorized self-attention (TSA) Mechanism. 66

4.1 Proposed **M**ulti-t**a**sk **S**emantic **P**arsing (MaSP) model. Note that P* and T* are predicate and entity type ids in Wikidata where entity type id originally starts with Q but is replaced with T for clear demonstration. 83

4.2 Transformation from entity-pointed logical form to KB-executable logical form for KB querying. 87

4.3 Performance of entity linking. “w/o ET” denotes removing entity type filtering. 95

4.4 An overview of the proposed two-objective learning based **S**emantic **T**riple **E**ncoder for **L**ink **P**rediction (STELP). This illustration is based on a corruption of tail entity, and in the same way for the corruption of head entity or even relation. Note that a notation whose superscript includes “ $'$ ” denotes it is derived from a negative example, otherwise from a positive one. 101

4.5 Taxonomy of different approaches to integrating pre-trained LMs with knowledge graphs. 114

4.6 Dev loss in pre-training phase for different training or masking schemes. Note only entity-level masking has $\mathcal{L}_R$. 130

Abbreviation

Natural Language Processing - NLP

Artificial Intelligence - AI

Part-of-Speech - POS

Named Entity Recognition - NER

Term Frequency–Inverse Document Frequency - TF-IDF

Deep Learning - DL

Deep Neural Network - DNN

Continuous Bag-of-Words - CBoW

Recurrent Neural Network - RNN

Long Short-Term Memory - LSTM

Gated Recurrent Unit - GRU

Convolutional Neural Network - CNN

Self-Attention Network - SAN

Multi-Layer Perceptron - MLP

Neural Machine Translation - NMT

Natural Language Inference - NLI

Language Modeling - LM

Masked Language Modeling - MLM

Next Sentence Prediction - NSP

Sentence Order Prediction - SOP

Span-Boundary Objective - SBO

Byte Pair Encoding - BPE

Out-of-Vocabulary - OOV

Feed-Forward Network - FFN

One-Dimensional Convolutional Neural Network - 1D-CNN

Maximum Likelihood Estimation - MLE

Bidirectional Encoder Representations from the Transformer - BERT

Knowlebidge Base - KB

Knowledge Graph - KG

Nomenclature and Notation

Numbers and Arrays

a	A scalar (integer or real)
$\boldsymbol{a}$	A vector
$\boldsymbol{A}$	A matrix
$\mathsf{\mathbf{A}}$	A tensor
$\boldsymbol{I}_n$	Identity matrix with n rows and n columns
$\boldsymbol{I}$	Identity matrix with dimensionality implied by context
$\boldsymbol{e}^{(i)}$	Standard basis vector $[0, \ldots, 0, 1, 0, \ldots, 0]$ with a 1 at position i
$\text{diag}(\boldsymbol{a})$	A square, diagonal matrix with diagonal entries given by $\boldsymbol{a}$
a	A scalar random variable
$\mathbf{a}$	A vector-valued random variable
$\mathbf{A}$	A matrix-valued random variable

Sets and Graphs

$\mathbb{A}$	A set
$\mathbb{R}$	The set of real numbers
$\{0, 1\}$	The set containing 0 and 1
$\{0, 1, \ldots, n\}$	The set of all integers between 0 and n
$[a, b]$	The real interval including a and b
$(a, b]$	The real interval excluding a but including b
$\mathbb{A} \backslash \mathbb{B}$	Set subtraction, i.e., the set containing the elements of $\mathbb{A}$ that are not in $\mathbb{B}$

$\mathcal{G}$	A graph
$Pa_{\mathcal{G}}(\mathrm{x}_i)$	The parents of x_i in $\mathcal{G}$

Indexing

a_i	Element i of vector $\boldsymbol{a}$, with indexing starting at 1
a_{-i}	All elements of vector $\boldsymbol{a}$ except for element i
$A_{i,j}$	Element i, j of matrix $\boldsymbol{A}$
$\boldsymbol{A}_{i,:}$	Row i of matrix $\boldsymbol{A}$
$\boldsymbol{A}_{:,i}$	Column i of matrix $\boldsymbol{A}$
$\mathsf{A}_{i,j,k}$	Element (i, j, k) of a 3-D tensor $\mathbf{\mathsf{A}}$
$\mathbf{\mathsf{A}}_{:,:,i}$	2-D slice of a 3-D tensor
a_i	Element i of the random vector $\mathbf{a}$

Calculus

$\dfrac{dy}{dx}$	Derivative of y with respect to x
$\dfrac{\partial y}{\partial x}$	Partial derivative of y with respect to x
$\nabla_{\boldsymbol{x}} y$	Gradient of y with respect to $\boldsymbol{x}$
$\nabla_{\boldsymbol{X}} y$	Matrix derivatives of y with respect to $\boldsymbol{X}$
$\nabla_{\mathbf{\mathsf{X}}} y$	Tensor containing derivatives of y with respect to $\mathbf{\mathsf{X}}$
$\dfrac{\partial f}{\partial \boldsymbol{x}}$	Jacobian matrix $\boldsymbol{J} \in \mathbb{R}^{m \times n}$ of $f : \mathbb{R}^n \to \mathbb{R}^m$
$\nabla_{\boldsymbol{x}}^2 f(\boldsymbol{x})$ or $\boldsymbol{H}(f)(\boldsymbol{x})$	The Hessian matrix of f at input point $\boldsymbol{x}$
$\int f(\boldsymbol{x}) d\boldsymbol{x}$	Definite integral over the entire domain of $\boldsymbol{x}$
$\int_{\mathbb{S}} f(\boldsymbol{x}) d\boldsymbol{x}$	Definite integral with respect to $\boldsymbol{x}$ over the set $\mathbb{S}$

Probability and Information Theory

$P(\mathrm{a})$	A probability distribution over a discrete variable
$p(\mathrm{a})$	A probability distribution over a continuous variable, or over a variable whose type has not been specified
$\mathrm{a} \sim P$	Random variable a has distribution P
$\mathbb{E}_{\mathrm{x} \sim P}[f(x)]$ or $\mathbb{E} f(x)$	Expectation of $f(x)$ with respect to $P(\mathrm{x})$
$\mathrm{Var}(f(x))$	Variance of $f(x)$ under $P(\mathrm{x})$
$\mathrm{Cov}(f(x), g(x))$	Covariance of $f(x)$ and $g(x)$ under $P(\mathrm{x})$
$H(\mathrm{x})$	Shannon entropy of the random variable x

$D_{\mathrm{KL}}(P\|Q)$	Kullback-Leibler divergence of P and Q
$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution over $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$

Functions

$f : \mathbb{A} \rightarrow \mathbb{B}$	The function f with domain $\mathbb{A}$ and range $\mathbb{B}$
$f \circ g$	Composition of the functions f and g
$f(\boldsymbol{x}; \boldsymbol{\theta})$	A function of $\boldsymbol{x}$ parametrized by $\boldsymbol{\theta}$. (Sometimes we write $f(\boldsymbol{x})$ and omit the argument $\boldsymbol{\theta}$ to lighten notation)
$\log x$	Natural logarithm of x
$\sigma(x)$	Logistic sigmoid, $\dfrac{1}{1+\exp(-x)}$
$\zeta(x)$	Softplus, $\log(1+\exp(x))$
$\|\|\boldsymbol{x}\|\|_p$	L^p norm of $\boldsymbol{x}$
$\|\|\boldsymbol{x}\|\|$	L^2 norm of $\boldsymbol{x}$
x^+	Positive part of x, i.e., $\max(0, x)$
$\mathbf{1}_{\mathrm{condition}}$	is 1 if the condition is true, 0 otherwise